COL 226: Programming Languages
Assignment2: Lexing and Parsing

Submission Lifeline: Wed 10 Mar 2021, 11:59 PM
Submission Deadline with Late Penalty: Sun 14 Mar 2021, 11:59 PM

Problem Statement

Your task is to write lexer and parser for Boolean algebra using ML-Lex and ML-Yacc.

Problem Specification

Create lexer and parser files using ML-Lex and ML-Yacc for SML. Add a make-file to generate executable named `a2`.

Input: The executable should take the name of the file to be analyzed as a command-line argument. The syntax of the program, along with the names of token types for each token, is as follows:

1. The input file consists of a single input program. Define a non-terminal `program` for the entire program. Each file should end with a token type `EOF`.
2. A program is a set of statements. A statement is a boolean formula followed by a 'termination character semicolon (";"). Use non-terminal `statement` for a valid statement and token type `TERM` for semicolon.
3. Represent a formula by non-terminal `formula`. A formula may consists of:
   (a) Constants "TRUE" and "FALSE" representing bool 1 and 0 respectively. Use token type `CONST` for constants.
   (b) Right-to-left associative prefix unary operator "NOT" of a formula, having form "NOT formula". Use token type `NOT` for token "NOT".
   (c) Left-to-right associative infix binary operators over two formulas, having form "formula1 binop formula2", where "binop" can be "AND", "OR", "XOR" and "EQUALS". Represent these using token types `AND, OR, XOR` and `EQUALS` respectively.
   (d) Right-to-left associative implication operator of the form "formula1 IMPLIES formula2". Use token type `IMPLIES` for token "IMPLIES".
   (e) Right-to-left associative if-then-else operator of the form "IF formula1 THEN formula2 ELSE formula3". Use token types `IF, THEN` and `ELSE` for "IF", "THEN" and "ELSE".
   (f) Parenthesis to define order of evaluation over different operations "(formula)". Use token types `LPAREN` and `RPAREN` for left and right parenthesis respectively.

(g) For formulas and sub-formulas without parenthesis, the order of evaluation is decided according to associativity rules.

(h) Any other string containing only lower and upper case English alphabets is a variable. Use token type `ID` for variable identifiers.

(i) All the operations mentioned in any point above have the same precedence, and the precedence is decreasing from point 3b to point 3e. For example, NOT operation has higher precedence than AND, OR, XOR and EQUALS operations.

4. A formula or statement may be written in several lines.

Executing the Program: We will compile your submission by running `make` command and run the executable as `./a2` ⟨`filename`⟩

Output: The executable `a2` should produce the output of the lexer followed by a newline, then the parser's output. Lexer output should be a comma-separated list (enclosed in square brackets) of tokens in order of their appearance in the input file. Each token in output should be of the form "⟨`token type`⟩ `space` ⟨`actual token in the input file enclosed in double quotes`⟩". The output of the parser should be the ~~preorder~~ POSTORDER traversal of the generated parse tree. The preorder traversal should be a comma-separated list of each node's representation. Use the production rules to represent a non-terminal node, and "⟨`token type`⟩ `space` ⟨`actual token in the input file`⟩" to represent the terminal nodes in preorder traversal.

Error Reporting: Whenever an invalid token is encountered, lexer should generate `Unknown Token:`⟨`line no`⟩`:`⟨`column number`⟩`:`⟨`token`⟩ error. Here ⟨`line no`⟩ and ⟨`column number`⟩ start from 1, and ⟨`token`⟩ is the invalid token. If the input is not syntatically correct according to the specifications, the parser should generate `Syntax Error:`⟨`line no`⟩`:`⟨`column number`⟩`:`⟨`production rule`⟩ error. Here ⟨`production rule`⟩ is the production rule where syntax did not match.

Examples: Some examples of valid and invalid programs are as follows:

1. `(xyz IMPLIES FALSE) OR TRUE AND IF A THEN b ELSE c;` is a valid statement having identifiers `xyz, A, b` and `c`. The lexer output for this example should be as follows:

   [`LPAREN` "(", `ID` "xyz", `IMPLIES` "IMPLIES", `CONST` "FALSE", `RPAREN`, ")", `OR` "OR", `CONST` "TRUE", `AND` "AND", `IF` "IF", `ID` "A", `THEN` "THEN", `ID` "b", `ELSE` "c", `TERM` ";"]

2. If line 5 of input file is `a | b;` lexer should result in error `Unknown token:5:3:|`.

3. `IF x EQUALS y z THEN TRUE ELSE a XOR b;` is an invalid statement. Since all the tokens in this expression are valid, the lexer should produce the correct output. However, the parser should generate an error. If this statement is in line 5, the parser error should be

   `Syntax Error:5:15:‘‘concerned production rule’’`

Tasks to be done

Task1 (20 marks): Write EBNF for the language given in this document in a file named `ebnf.txt`. Take care of the order of evaluation specified here.
Task2 (30 marks): Create a file for lexical analysis using ML-Lex for the language given in this document. Follow the specifications given regarding input, output, and execution.
Task3 (50 marks): Create a file for parsing the language given in this document using ML-Yacc. Follow the specifications given regarding input, output, and execution.

Submission Instruction

Submit a zip file named ⟨EntryNumber⟩.zip. On unzipping the file, it should produce the lexer, parser, `ebnf.txt` and `makefile`. It should also contain any other source file used in your program. You may provide a README.md that contains suggestions for the evaluator.

Important Notes

1. Do not change any of the names given in this document. You are not even allowed to change upper-case letters to lower-case letters or vice-versa.

2. Follow the input/output specification as given. A part of this assignment will be auto-grade. In case of mismatch, you will be awarded zero marks.

3. Take care of extra whitespace characters.

4. You may create new token types if required. Make sure that all the terminals use the token types specified in the document.

5. We have created a piazza post titled "A2 Queries". If you have doubts, please post only in this thread. The queries outside this thread or over email *will not* be entertained.